GLM: General Language Model Pretraining with Autoregressive Blank Infilling

Abstract

There have been various types of pretraining architectures including autoencoding models (e.g., BERT), autoregressive models (e.g., GPT), and encoder-decoder models (e.g., T5). However, none of the pretraining frameworks performs the best for all tasks of three main categories including natural language understanding (NLU), unconditional generation, and conditional generation. We propose a General Language Model (GLM) based on autoregressive blank infilling to address this challenge. GLM improves blank filling pretraining by adding 2D positional encodings and allowing an arbitrary order to predict spans, which results in performance gains over BERT and T5 on NLU tasks. Meanwhile, GLM can be pretrained for different types of tasks by varying the number and lengths of blanks. On a wide range of tasks across NLU, conditional and unconditional generation, GLM outperforms BERT, T5, and GPT given the same model sizes and data, and achieves the best performance from a single pretrained model with 1.25× parameters of BERT Large , demonstrating its generalizability to different downstream tasks.